1-2019

Fixed Choice Design and Augmented Fixed Choice Design for Network Data with Missing Observations

Miles Q. Ott
Smith College, mott@smith.edu

Matthew T. Harrison
Brown University

Krista J. Gile
University of Massachusetts Amherst

Nancy P. Barnett
Brown University

Joseph W. Hogan
Brown University

Biostatistics (2019) **20**, 1, *pp.* 97–110
doi:10.1093/biostatistics/kxx066
Advance Access publication on December 16, 2017

Fixed choice design and augmented fixed choice design for network data with missing observations

MILES Q. OTT*

Statistical and Data Sciences Program, Smith College, 7 College Lane, Northampton, MA, 01063 USA

mott@smith.edu

MATTHEW T. HARRISON

Division of Applied Mathematics, Brown University, 170 Hope St, Providence, RI 02906, USA

KRISTA J. GILE

Department of Mathematics and Statistics, University of Massachusetts Amherst, 710 N. Pleasant Street, Amherst, MA 01003-9305, USA

NANCY P. BARNETT

Department of Behavioral and Social Sciences, Brown University School of Public Health, Box G-S121-4, Providence, RI 02912, USA

JOSEPH W. HOGAN

Department of Biostatistics, Brown University School of Public Health, 121 South Main Street, Providence, Rhode Island 02903, USA

SUMMARY

The statistical analysis of social networks is increasingly used to understand social processes and patterns. The association between social relationships and individual behaviors is of particular interest to sociologists, psychologists, and public health researchers. Several recent network studies make use of the fixed choice design (FCD), which induces missing edges in the network data. Because of the complex dependence structure inherent in networks, missing data can pose very difficult problems for valid statistical inference. In this article, we introduce novel methods for accounting for the FCD censoring and introduce a new survey design, which we call the augmented fixed choice design (AFCD). The AFCD adds considerable information to analyses without unduly burdening the survey respondent, resulting in improvements over the FCD, and other existing estimators. We demonstrate this new method through simulation studies and an analysis of alcohol use in a network of undergraduate students living in a residence hall.

Keywords: Augmented fixed choice design; Fixed choice design; Missing data; Right censoring by degree; Social network.

*To whom correspondence should be addressed.

1. Introduction

The statistical analysis of social networks is increasingly used to understand social processes and patterns (Knoke and Yang, 2008). Of particular interest to sociologists, psychologists, and public health researchers is the association between social relationships and individual behaviors. Social relationships are often measured through nominations, such as when collecting information on a friend network, the nomination from person i to person j indicates that person i claims person j as their friend. When analyzing a network, there is often the possibility that nominations are missing (Kossinets, 2006). Because of the complex dependence structure inherent in networks, missing data can pose very difficult problems (Holland and Leinhardt, 1973; Marsden, 1990; Kossinets, 2006). Even when missingness is at random, it can induce bias in structural measures of the network, such as homophily and centrality (Smith and Moody, 2013). Of particular interest is how to handle missingness when analyzing the network for peer effects on behavior.

A considerable amount missingness is a result of study design. Several recent studies on networks in school settings make use of the fixed choice design (FCD), which typically induces missing edges. In FCD, the number of possible nominations that each person in the network can make is capped at a maximum, m, inducing missing nominations (Holland and Leinhardt, 1973; Kossinets, 2006; Yan and Gregory, 2011). For example, studies have variously restricted the number of friends in a classroom to 4 when studying depression (Witvliet *and others*, 2010), and the number of best friends to 5 when studying smoking (Mercken *and others*, 2010), and another study on infectious disease transmission on social networks allowed participants to name up to 6 within class contacts and up to 4 outside of class contacts (Conlan *and others*, 2010). The National Longitudinal Study of Adolescent Health (Add Health) allowed participants to nominate and rank up to 5 boys and 5 girls as friends (Resnick *and others*, 1997; Goodreau, 2007) and is the basis for significant methodological advancements in the analysis of social networks with missing data. For example, Goodreau *and others* (2009) approach the Add Health censoring mechanism by assuming that the true network of interest is the network consisting of the top five male and top five female friends of each respondent. Hipp *and others* (2015) investigated the consequences of missing observations in longitudinal network data and found that different methods of accounting for missingness can lead to vastly different results. Wang *and others* (2016) present an exponential random graph (ERGM) method for the imputation of missing network data using the Add Health study as an application. Handcock and Gile present network modeling approaches for networks with missing data with application to the Add Health study (Handcock and Gile, 2007).

Holland and Leinhardt (1973) first introduced the problem of missing ties due to the FCD (also called limited choice design and right-censoring of degree). Kossinets (2006) subsequently showed how the FCD can lead to biases in estimates of structural measures of the network. Gommans and Cillessen (2015) compared analyses on the same populations of elementary school students, with FCD as well as a design without censoring, and found significant differences in the conclusions drawn from the two different data collection schemes.

Hoff *and others* (2013) have developed a likelihood based approach for fixed rank network data (where there is a maximum number of nominations that could be made, and those nominations are ranked) as well as for FCD, and then used these likelihoods in Bayesian estimation of latent variables which are assumed to govern the nominations and their ranks.

In a study on the transmission of influenza through household contacts, Mossong *and others* (2008) collected egocentric information on the number of household contacts an individual in the household has made in a given day, without collecting which specific household members the ego was in contact with. Potter *and others* (2011) used these data to model disease transmission between household members, showing that the number of contacts, or edges, can be useful information even when m is capped at zero.

In this work, we introduce a novel approach that can improve inference for FCD data: that true total nominations are collected in addition to the standard FCD data. We develop a method that accounts for

the missingness resulting from FCD, given the true total nominations, and show how different censoring cut-offs affect parameter estimation using a dyad-independent ERGM model. In Section 2, we introduce novel methods for accounting for the fixed choice censoring and introduce a new survey design. In Section 3, we present simulation studies wherein we demonstrate our methods for handling fixed choice data and compare the effects of different censoring cut-offs on parameter estimation, as well as against different estimation methods. In Section 4, we demonstrate our method in an analysis of relationships in the presence of alcohol use in a network of undergraduates living in a residence hall. A discussion is presented in Section 5.

2. Model formulation and inference

2.1. *General framework*

Given a family of probability models indexed by a parameter β we use the notation $p_\beta(z) = \mathbb{P}(Z = z \mid \beta)$ to denote the probability mass function (pmf) of a discrete random variable Z under the model with parameter β. Usually, the model will also involve fixed covariates, but this is suppressed for now in our notation. Our goal is to identify what $p_\beta(z)$ is so that we can find the maximum likelihood estimates of β given our data z. If $z = (z_{ij})$ is a matrix, then we use z_i to denote the ith row of z. If z is a vector, then we use $s(z) = \sum_i z_i$ to denote the sum of z.

Let Y be an $n \times n$ sociomatrix, where $Y_{ij} = 1$ if i nominates j and $Y_{ii} = 0$ for all i. Note that $Y_{ij} = 1$ does not imply that $Y_{ji} = 1$, meaning that these relationships may be unreciprocated. We use $R = (R_1, \ldots, R_n)$ to denote the row sums of Y, i.e., $R_i = s(Y_i)$. We assume that Y is the sociomatrix that would be observed without any reporting constraints, and, hence, R_i is the true total of nominations of the ith subject. If Y could be observed, then, given a computationally tractable probability model $p_\beta(y)$, we could use standard likelihood-based methods to estimate β. A simple model, which we assume here, is that all ties are independent Bernoulli random variables, namely,

$$p_\beta(y) = \textstyle\prod_{ij} \pi_{ij}(\beta)^{y_{ij}} (1 - \pi_{ij}(\beta))^{1-y_{ij}}, \tag{2.1}$$

where $\pi_{ij}(\beta) = \mathbb{P}(Y_{ij} = 1 \mid \beta)$ is designed according to the problem at hand.

In a FCD that allows at most m nominations per subject we do not observe Y, but instead observe a censored sociomatrix W. We assume the following possible censoring mechanism: if $R_i \le m$, then there is no censoring and $W_i = Y_i$, otherwise, the subject reports exactly m of the R_i original nominations chosen uniformly at random. Consequently, the joint pmf of (W, R) is

$$p_\beta(w, r) = \textstyle\prod_i p_\beta(w_i, r_i) \tag{2.2}$$

with

$$\begin{aligned} p_\beta(w_i, r_i) &= \begin{cases} \mathbb{P}(Y_i = w_i \mid \beta) & s(w_i) = r_i \le m \\ \binom{r_i}{m}^{-1} \mathbb{P}(Y_{ij} = 1 \text{ for all } j \text{ with } w_{ij} = 1, \text{ and } R_i = r_i \mid \beta) & s(w_i) = m < r_i \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \prod_j \pi_{ij}(\beta)^{w_{ij}} (1 - \pi_{ij}(\beta))^{1-w_{ij}} & s(w_i) = r_i \le m \\ \binom{r_i}{m}^{-1} \prod_{j:w_{ij}=1} \pi_{ij}(\beta)\, \mathbb{P}\left(\sum_{j:w_{ij}=0} Y_{ij} = r_i - m \,\middle|\, \beta \right) & s(w_i) = m < r_i \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.3}$$

Equation (2.3) involves the term

$$\mathbb{P}\left(\sum_{j:w_{ij}=0} Y_{ij} = r_i - m \,\middle|\, \beta\right), \tag{2.4}$$

which is simply the probability that a sum of independent Bernoulli's is equal to $r_i - m$ and is easy to compute using discrete convolution.

From (2.2) to (2.4), we see that $p_\beta(w, r)$ is easily computable and can be combined with standard likelihood methods to generate estimates of β from joint observations of W and R. Since R is not usually observed in a FCD, we call this design an augmented FCD (AFCD). In a regular FCD, from (2.2) we also see that

$$p_\beta(w) = \textstyle\prod_i p_\beta(w_i), \tag{2.5}$$

where we sum over all possible values of r_i for rows i with missing data:

$$p_\beta(w_i) = \begin{cases} \prod_j \pi_{ij}(\beta)^{w_{ij}}(1-\pi_{ij}(\beta))^{1-w_{ij}} & s(w_i) = r_i < m \\ \sum_{r_i=m}^{n-1} \binom{r_i}{m}^{-1} \prod_{j:w_{ij}=1} \pi_{ij}(\beta)\, \mathbb{P}\left(\sum_{j:w_{ij}=0} Y_{ij} = r_i - m \,\middle|\, \beta\right) & s(w_i) = m \le r_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

Note from (2.3) when we are finding the $p_\beta(w_i, r_i)$, we are in the fully observed data case when $r_i \le m$, which is in contrast to (2.6) when there are missing observations $r_i = m$ since we do not observe r_i. This is because when r_i is known, then we are able to discern whether we have complete data when $s(w_i) = m$. However when r_i is censored, as in the FCD setting, we cannot be certain if $s(w_i) = m = r_i$ or if $s(w_i) = m < r_i$.

As noted above, our goal for both the AFCD and FCD is to identify the function $p_\beta(w_i, r_i)$ in the AFCD setting, or the function $p_\beta(w_i)$ in the FCD setting so that we may find the maximum likelihood estimates of β given our data. Now that $p_\beta(w_i, r_i)$ and $p_\beta(w_i)$ have been specified for the AFCD and FCD data cases, respectively, we employ one of the many available optimization techniques to find the maximum likelihood estimates of of β given the available data.

2.2. *Variance estimation*

The variance estimation for $\hat{\beta}$ is non-trivial. Simply using the observed information will not incorporate the uncertainty of the censored values. In order to quantify the variance of the maximum likelihood estimates of the β parameters, we describe a parametric bootstrap by following these steps with B bootstrap samples.

1. Maximize the appropriate likelihood as described above [(2.3) for AFCD or (2.6) for FCD] to produce $\hat{\beta}$.
2. Using $\hat{\beta}$ and the same covariates used in the model in Step 1, generate sociomatrix $Y^b, b \in 1, 2, \ldots B$.
3. With uniform probability delete $r_i - m$ edges for all row i for i in $1, \ldots, n$ of Y^b to generate W_b.
4. Maximize the appropriate likelihood using W_b to attain $\widehat{\beta_b}$
5. Repeat Steps 2–4 B times to generate a distribution of $\widehat{\beta_b}$ which can be used to get bootstrapped standard errors and confidence intervals.

3. Simulation studies

Using the general framework described above, we experiment with models of the form

$$g(\pi_{ij}(\beta)) = \beta^{\mathrm{T}} x_{ij},$$

where g is an appropriate link function for binary data, and the parameter $\beta \in \mathbb{R}^d$ is a column vector and where $x_{ij} \in \mathbb{R}^d$ is a column vector of known covariates that may depend on both i and j. If Y was fully observed, so that we could use $p_\beta(y)$ from (2.1) for inference, then this would be regression with edge-level covariates. Instead, for a FCD or an AFCD, we use $p_\beta(w)$ or $p_\beta(w, r)$, respectively, as derived in the previous section to find a maximum likelihood estimates. We proceed in the rest of this article to use the probit link function, though other link functions could also be implemented.

We do not directly observe edge-level covariates in this simulation, but rather create them from vertex-level covariates. For each vertex i, let $v_i \in \mathbb{R}^q$ be a vector of known covariates. Define the edge-level covariates as some subset of

$$x_{ij} = (1, v_{i1}, \ldots, v_{iq}, v_{j1}, \ldots, v_{jq}, |v_{i1} - v_{j1}|, \ldots, |v_{iq} - v_{jq}|)$$

which has dimension $d = 3q + 1$. For example, if we look at age as the single variable of interest ($q = 1$), then we define:

$$x_{ij} = (1, \mathrm{Age}_i, \mathrm{Age}_j, |\mathrm{Age}_i - \mathrm{Age}_j|)$$

for all i and j. If we were to use age and income as the two covariates of interest ($q = 2$), then we define:

$$x_{ij} = (1, \mathrm{Age}_i, \mathrm{Income}_i, \mathrm{Age}_j, \mathrm{Income}_j, |\mathrm{Age}_i - \mathrm{Age}_j|, |\mathrm{Income}_i - \mathrm{Income}_j|)$$

for all i and j.

3.1. *Simulation design*

In order to contrast the AFCD, FCD, a naive analysis (where we assume that all unobserved values of W are non-edges), and Hoff *and others* (2013)'s censored binary (CB) estimator, we performed a simulation study.

For a simulated population of size n, we first generated a continuous covariate V for all n members of the simulated population, which stays fixed for all simulations. In keeping with the previously defined notation, we next generated directed edges between members of the network such that the probability that individual i has a directed relationship with individual j, denoted by $Y_{ij} = 1$, is independent given X_{ij}:

$$\mathrm{probit}(\mathbb{P}(Y_{ij} = 1 \mid \beta, x_{ij})) = \beta^{\mathrm{T}} x_{ij},$$

where

$$x_{ij}^{\mathrm{T}} = (1, v_i, |v_i - v_j|)$$

and β is a 3×1 vector. Note that our formulation allows for both row covariates (v_i), column covariates (v_j) as well as edge covariates ($|v_i - v_j|$), and that here we do not use the column covariates. We next simulate the censoring processes of the AFCD, the FCD, and the CB (which has the identical censoring process as the FCD) for maximum number of nominations $m \in \{0, 2, 4, 6, 8\}$. Under the AFCD, in row

i, given r_i and m, when $r_i > m$, $r_i - m$ edges are censored, where each of the r_i edges have an equal probability of being censored. Moreover, all non-edges in rows where $r_i > m$ are censored. This gives us the observed AFCD data W_{AFCD}. For the FCD, for a maximum m in each row where $r_i \geq m$, $r_i - m$ edges are censored, and all non-edges in these rows are censored, to give us the observed FCD data W_{FCD}.

We then obtain maximum likelihood estimators under the AFCD and FCD with varying m values. We also find estimates for the CB using the posterior means of β using the `amen` package in R (Hoff *and others*, 2015). Finally, we carry out the naive analysis by applying probit regression to the W_{FCD} data where we treat all censored values as zero, which is the default current practice for analyzing FCD.

3.2. *Simulation results*

The distribution of the covariate value V is displayed in Figure 1(a). We generated the network using $\beta_0 = -1, \beta_1 = 0.02, \beta_2 = -0.025$ which generates networks with a roughly normal distribution of edges with mean number of edges = 10; however, having a normal distribution of edges is not a requirement here. Here, beta $\beta_0 = -1$ implies that for i,j where $v_i = v_j = 0$, the probit of i nominating j is equal to -1. Likewise, $\beta_1 = 0.02$ implies that for a fixed value in the absolute difference between v_i and v_j, as v_i increases by one unit, then the probit of i nominating j increases by 0.02. Last, $\beta_2 = -0.025$ implies that for a fixed value of v_i as the absolute difference between v_i and v_j increases by one unit, the probit of i nominating j decreases by 0.025. We simulated 100 different networks with 100 nodes using these covariate and β values. The distribution of r_i for all 100 simulations is plotted in Figure 1(b), where grey bars indicate m values used to censor the simulated data ($m \in 0, 2, 4, 6, 8$). For the CB, which uses Bayesian inference, we used MCMC and generated 55 000 posterior draws. The first 5000 draws were discarded, and of the remaining 50 000 posterior draws, we used every 25th draw, for a total of 2000 posterior draws. We present means of these 2000 posterior draws as the estimates.

We present the mean squared error (MSE), empirical bias, and SEs of the estimated β's from the 100 simulations in Table 1.

For a given maximum m, the AFCD had lower MSE than the FCD and CB for β_0, β_1, and β_2. The AFCD had lower MSE than the naive estimator (in which the data are falsely assumed to be uncensored) for β_0 and β_2, though naive estimator had lower MSE than the AFCD for β_1, due to its lower variance. For the β_0 parameter, having an AFCD with $m = 0$ had a lower MSE than FCD with $m = 2$, naive with $m = 2$ and CB with $m = 6$. For the β_1 parameter, having an AFCD with $m = 0$ outperformed a FCD with $m = 4$ and CB with $m = 6$ in terms of MSE. For β_2, the AFCD had lower MSE for all levels of m simulated as compared to the other estimators. For higher values of m, the relative improvement in MSE for AFCD over FCD tended to diminish. In other words, when the data collection design incurs a higher levels of missingness (such as when the maximum number of nominations that can be made in the survey m is low relative to r_i, the true total nominations) the added benefit of knowing the true number of relationships can be quite large, which is why the AFCD will outperform the FCD and CB. When m is high relative to the distribution of the true total nominations per individual in the network, there will be less missingness and a smaller benefit of the AFCD as compared to the FCD and CB. In general, the AFCD requires lower m to achieve comparable MSE to the other estimators, indicating that it may be preferable to collect the total number of relationships r to marginally increasing m.

When comparing the MSE of the FCD to the CB, neither estimator proved uniformly better. While the FCD had substantially lower MSE than the CB when estimating β_0 regardless of m, when estimating β_1 and β_2 the CB has lower MSE when m is smaller, and the FCD has lower MSE when m is larger.

The naive analyses in which all censored edges are considered to be non-edges (as is a common current practice), estimated β_0 poorly. This is result is not surprising as the naive estimator will necessarily underestimate the probability of any edge due to the way that it treats all censored values as non-edges. Notably, the naive estimator tended to have high bias and a low standard deviation.

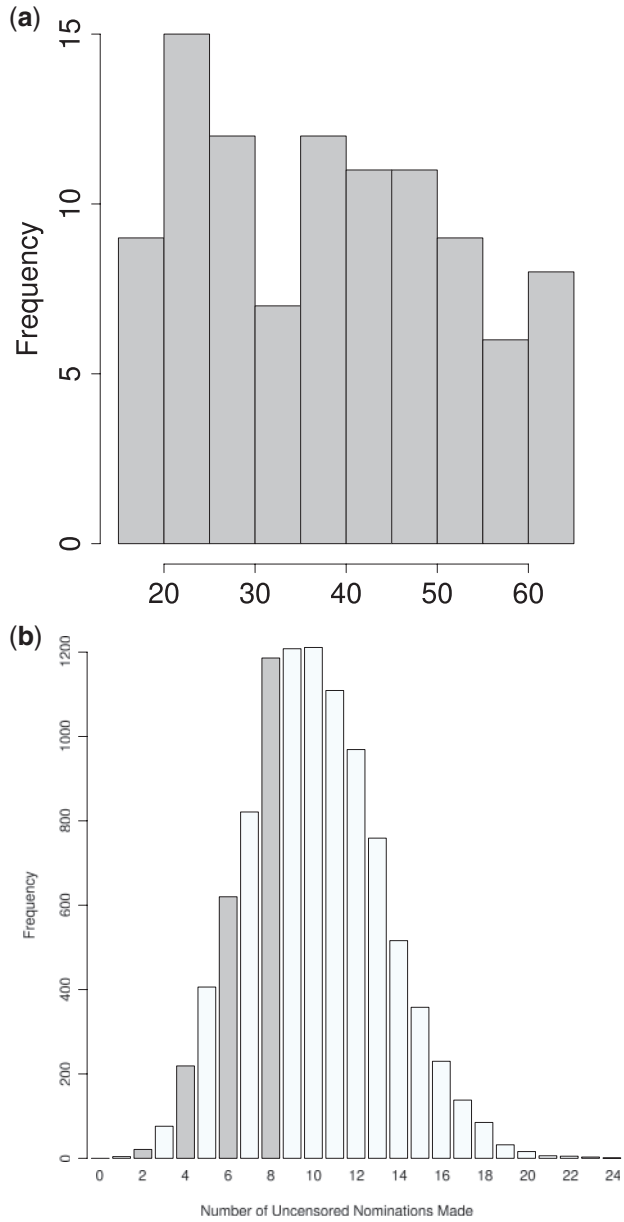

Fig. 1. Simulation specifications: distribution of covariates used to generate simulated networks, (a), and (b) distribution of the true total nominations made from 100 simulations. The grey bars denote values of m used to impose censoring in 100 different simulations ($m \in 0, 2, 4, 6, 8$).

4. Analysis of UrWeb study

We next apply the method to the UrWeb data set, in which data were collected from residents of a primarily freshman dormitory (Barnett *and others*, 2014). Each participant was 18 years old or older when the survey was administered and was asked to report the number of days in a month that they consumed alcohol. Central to our interests here, each participant was asked to nominate which of the other

Table 1. *Mean squared error, empirical bias, and standard deviation of estimated β_0, β_1, and β_2 from 100 simulations, varying the maximum number of nominations observed, m*

m	β_0 MSE $\times 10^2$				β_0 |Bias| $\times$ 10				β_0 SD $\times 10$			
	AFCD	FCD	Naive	CB	AFCD	FCD	Naive	CB	AFCD	FCD	Naive	CB
0	34.68	—	—	—	4.43	—	—	—	3.90	—	—	—
2	0.34	153.08	73.84	377.03	0.01	7.85	8.59	18.84	0.59	9.61	0.26	4.72
4	0.27	32.00	28.31	501.85	0.02	2.46	5.32	18.12	0.52	5.12	0.20	13.24
6	0.26	1.10	10.42	77.28	0.03	0.09	3.22	7.20	0.51	1.05	0.23	5.07
8	0.25	0.40	3.26	4.95	0.03	0.06	1.79	1.77	0.50	0.64	0.28	1.36
m	β_1 MSE $\times 10^5$				β_1 |Bias| $\times 10^4$				β_1 SD $\times 10^2$			
	AFCD	FCD	Naive	CB	AFCD	FCD	Naive	CB	AFCD	FCD	Naive	CB
0	5.13	—	—	—	48.38	—	—	—	0.53	—	—	—
2	0.17	27.22	0.04	16.88	1.30	18.21	5.32	16.51	0.13	1.65	0.03	1.30
4	0.16	10.85	0.02	108.48	1.32	13.08	3.70	27.18	0.13	1.04	0.03	3.30
6	0.16	0.60	0.03	15.53	1.20	1.11	3.27	10.70	0.13	0.25	0.04	1.25
8	0.16	0.26	0.05	0.98	1.06	2.53	2.52	0.09	0.13	0.16	0.06	0.31
m	β_2 MSE $\times 10^5$				β_2 |Bias| $\times 10^4$				β_2 SD $\times 10^3$			
	AFCD	FCD	Naive	CB	AFCD	FCD	Naive	CB	AFCD	FCD	Naive	CB
0	936.23	—	—	—	651.87	—	—	—	71.89	—	—	—
2	1.30	45.91	5.95	3.54	3.45	134.78	71.59	50.70	3.61	16.74	2.89	3.13
4	0.58	6.14	2.97	1.48	2.32	39.16	50.54	31.84	2.40	6.82	2.06	2.18
6	0.37	0.43	1.54	0.63	1.10	2.09	35.04	17.30	1.93	2.07	1.77	1.84
8	0.31	0.33	0.78	0.34	0.29	0.88	22.23	6.23	1.76	1.84	1.70	1.74

participants were important to them. This network is pictured in Figure 2(a). Among the 129 participants included in the sample, 507 nominations were made; 4 participants did not nominate anyone nor were they nominated.

The UrWeb data were collected under a FCD, with $m = 10$. In this data set, only one person endorsed the maximum number of nominations, which suggests that there is little design-induced missingness of nominations. We will proceed to show the utility of the methods introduced here by artificially inducing a $m < 10$ in the UrWeb data set, estimating parameters, and comparing the estimated parameters when $m \in \{0, 2, 4, 6, 8\}$ to the parameters estimated with probit regression using the full data set. We will artificially induce $m \leq 8$ by deleting edges of individuals with more than m in two different ways, first by randomly deleting edges with uniform probability (as above in the simulation study), and secondly by deleting edges with regard to the order that each individual made their nominations. Figure 2(b) displays the distribution of the number of nominations that were made by each participant in the UrWeb study. We will proceed assuming that the UrWeb data are fully observed, and will demonstrate how this method will work in practice, compare the information loss for different m values, and contrast AFCD, FCD, CB, and naive analyses.

We use the number of days in a month that the subjects consume alcohol as the v covariate in this model:

$$\text{probit}(\mathbb{P}(Y_{ij} = 1|\beta, x_{ij})) = \beta^{\mathrm{T}} x_{ij},$$

where $Y_{ij} = 1$ indicates that participant i nominated j, and $x_{ij}^{\mathrm{T}} = 1, v_i, |v_i - v_j|$.

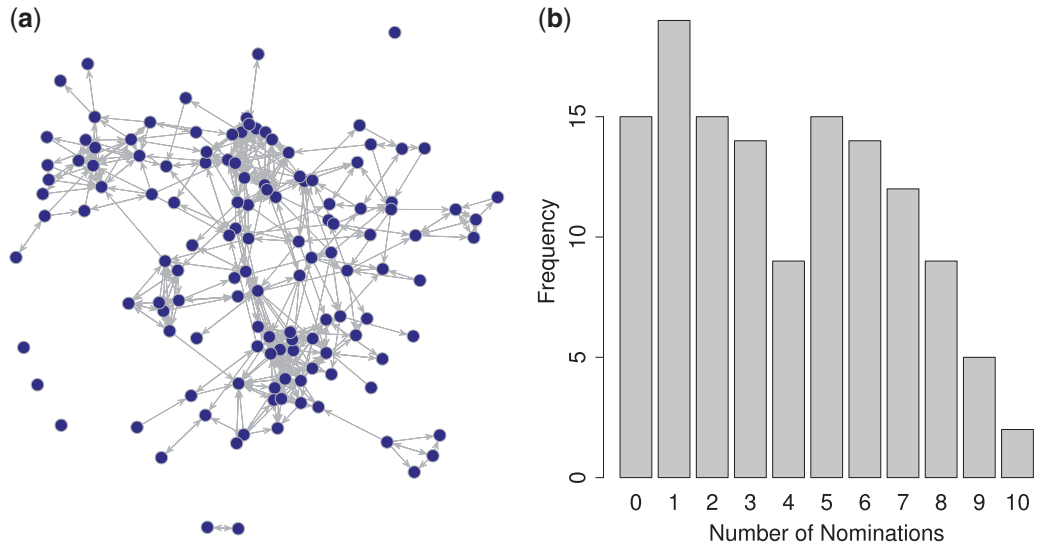

Fig. 2. UrWeb network (a), and distribution (b) distribution of nominations made by UrWeb study participants.

Having artificially induced $m \in \{0, 2, 4, 6, 8\}$ 100 times, we estimated $\beta_0, \beta_1, \beta_2$ using the AFCD, FCD, CB, and naive methods. We present boxplots of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ in Figure 3. In each of these figures, we denote the β estimates from the fully observed UrWeb data (where $m = 10$) with a solid horizontal line and the β estimate from the full data $\pm$ the estimated standard error from the full data with dotted horizontal lines.

When $m = 0$, we are only able to obtain maximum likelihood estimates of β for the AFCD. Because there is only one way to impose censoring when no nomination data is observed, Figure 3 presents a point estimate rather than a distribution of estimates when $m = 0$.

The analyses in this section differ from those in Section 3 in a few important ways. First, rather than simulating several networks, we are analyzing a single real network Y_{UrWeb} and repeatedly removing edges at random to form many different realizations of W_{UrWeb}. In these analyses, we do not know the true β values, and so we cannot evaluate the bias of the estimates. However, we can use these analyses to investigate information loss due to the design-induced censoring by comparing AFCD, FCD, naive, and CB estimates when $m < 10$ to estimates when we observe $m = 10$. For example in Figure 3, excluding when $m = 0$ the estimates of β_0 and β_1 from the AFCD are all within one standard error of the estimate when the full UrWeb data are observed ($m = 10$). This is in sharp contrast to the FCD, naive, and CB estimates of β_0 and β_1 when $m = 2, 4$. In general, the AFCD seems to lose less information than the FCD, which in turn loses less information than the naive analysis and the CB.

In this analysis of the UrWeb network, the AFCD produces estimates of β that are roughly centered on the full data estimate for β. The FCD method produces estimates that diverge from the estimate when the data are fully observed, especially when m is small. This result is in agreement to the simulation study which also showed that the FCD on average produces somewhat biased estimates when m is small.

Next, we deleted nominations in the reverse order in which they were made by the participants in the UrWeb study so that $m \in (0, 2, 4, 6, 8)$. We estimated β using AFCD, FCD, CB, and naive methods. For the AFCD and FCD, we calculated standard error estimates from 500 bootstrap samples. For the naive analyses, we simply used the standard error from a regular probit regression model. For the CB, we used the standard deviation of the posterior draws of the β terms to estimate the uncertainty of the estimates. These are presented in Figure 4. The UrWeb study was not explicitly designed to accommodate the AFCD,

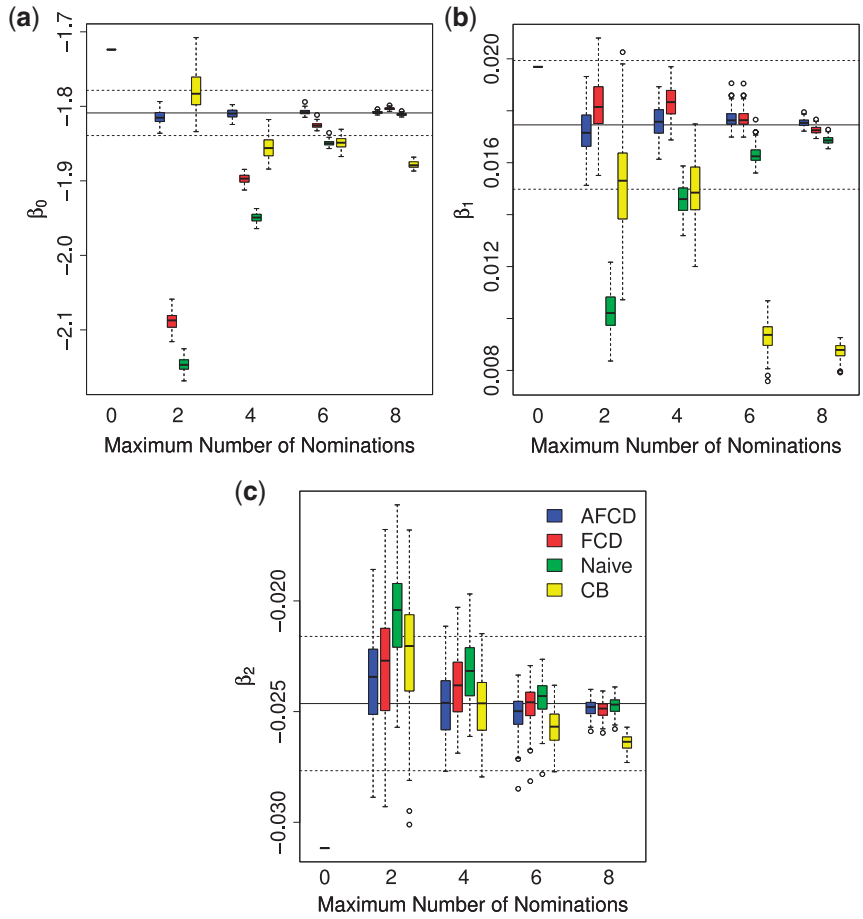

Fig. 3. Boxplots of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, varying maximum number of nominations m, and AFCD, FCD, naive analysis, and CB in the UrWeb data set. The black horizontal line is the estimated value of the β parameter when using the full UrWeb data. The dotted lines are $\beta \pm \widehat{\mathrm{SE}}(\beta)$ computed from the full UrWeb data for: (a) β_0, (b) β_1, and (c) β_2, respectively.

FCD, or CB design in that participants were not prompted to name a random sample of the people who were important to them. It is possible that study participants chose their nominations non-uniformly, for example nominating peers in the order of their importance. By deleting nominations in reverse order, we seek to investigate whether this could impact inference. We see very similar results when comparing Figure 3 in which nominations were deleted independently of order to Figure 4 in which nominations were deleted in reverse order. These results suggest that the order in which nominations were made in this data set did not greatly impact the inferences made in these analyses.

5. DISCUSSION

Collecting complete social network information in a closed population may be difficult as the network survey will impose an unreasonable amount of respondent burden. The FCD seeks to ameliorate respondent burden by asking respondents to nominate up to m individuals in the population with whom they have a particular relationship.

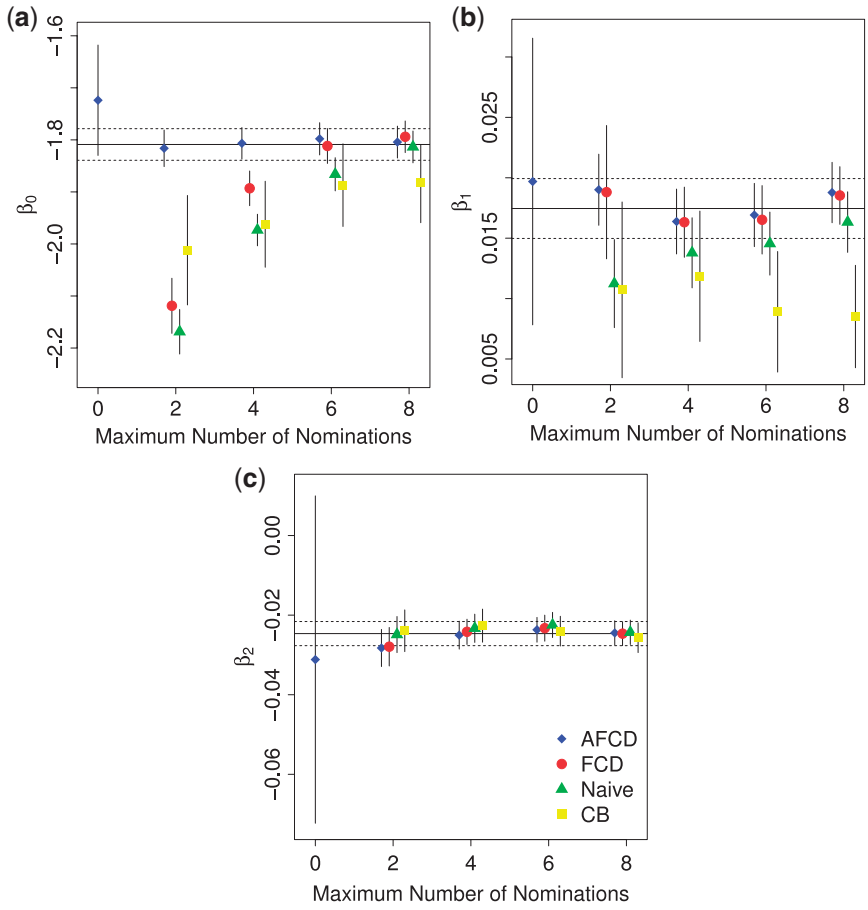

Fig. 4. Plots of $\hat{\beta} \pm 1$ bootstrap standard error for the AFCD and FCD, the probit regression standard error for the naive analysis, and the standard deviation of the posterior distribution estimates for the CB, deleting nominations in the reverse order in which they were made. The dotted lines are $\beta \pm \widehat{\mathrm{SE}}(\beta)$ computed from the full UrWeb data for: (a) β_0, (b) β_1, and (c) β_2, respectively.

In our application, we demonstrate that estimating associations between behaviors and social relationships from social network data arising from a fixed choice survey design as though the social network was fully observed (as is the current standard practice) can result in severely biased estimates. We introduce observed data likelihoods for FCD data. We demonstrated that maximizing the observed data likelihood for the FCD may improve the MSE in comparison to estimates where the data are (falsely) assumed to be fully observed.

We also introduce the AFCD, a new network survey sampling design and method of analysis which collects information on the total number of relationships for each individual in the network, in addition to the data collected with the standard FCD. This novel study design can add considerable information to analyses without unduly burdening the survey respondent, resulting in improvements over the FCD and naive analyses. We demonstrate that the AFCD is superior to both the FCD and naive analyses, as well as Hoff *and others* (2013)'s CB in terms of MSE. The improvement of the AFCD's MSE relative to the FCD, CB, and naive analyses is particularly pronounced when the m is small relative the number of true total

nominations. Unsurprisingly, our simulations show that for every estimator when m is larger, variation and bias is smaller. While collecting all nominations from each survey respondent would be optimal in terms of minimizing variance and bias, the AFCD can provide a way to improve estimation while keeping respondent burden low.

Since the AFCD utilizes information on the true total nominations, the estimates of the intercepts are much better with the AFCD than the FCD or the naive analyses. This suggests that the AFCD should be implemented when edge prediction is a goal of the analyses.

Limitations are acknowledged. In this work, we assume that nominations are randomly censored. Violation of this assumption may lead to incorrect inference. This assumption warrants additional investigation, and further research into survey methodology for AFCD and FCD data are necessary. Though, in this work we find that the order in which respondents nominated their peers did not heavily influence inference.

The analyses and simulations we present use a dyad independent ERGM model. This model does not incorporate important network characteristics including reciprocity, transitivity, and clustering. Modeling network structural characteristics and allowing for complex dependencies is particularly important when the goal of the model is to impute missing edges, or provide a realistic network model. Alternative models that incorporate network characteristics and dependencies include the social relations model (Warner *and others*, 1979), the ERGM family of models (Frank and Strauss, 1986; Robins *and others*, 2004; Goodreau, 2007), and the latent space and factor models (Hoff *and others*, 2002; Hoff, 2009).

Hoff *and others* (2013) presented likelihoods for fixed rank and FCD data. Hoff *et al.* assume that there is an underlying parametric model for the network that generates the ranked or binary social relations data. Using a social relations model, Hoff *et al.* perform estimation in the Bayesian framework. A benefit of that approach is the ability to accommodate both ranked and binary nominations. However, that method relies upon an underlying parametric model, requiring more stringent assumptions. As we are concerned with binary and not ranked data, we have compared the performance of Hoff's CB estimator to the AFCD, FCD, and naive estimator and found that in simulations the AFCD had uniformly lower MSE than the CB, while the FCD often had lower MSE than the CB. It should be noted that Hoff *et al.* found that their CB estimator performed comparably to their estimator that accounted for social rankings (Hoff *and others*, 2013), hence we would anticipate that the AFCD would also outperform the fixed rank estimator. Therefore we suggest that when collecting sociometric data, whether or not relationship rankings are collected, that the total number of relationships should be collected, so that censoring can be more readily accounted for.

Funding

NSF (SES-1230081) including support from the National Agricultural Statistics Service; NSF D(MS-1309004); Research Excellence Award from the Center for Alcohol and Addiction Studies, Brown University; and NIH (R01AA023522, R01CA183854, R01AI108441, P01AA019072, P30AI42853).

References

Barnett, N., Ott, M. Q., Rogers, M., Loxley, M., Linkletter, C. and Clark, M. (2014). Peer associations for substance use and exercise in a college student social network. *Health Psychology* **33**, 1134–1142.

Conlan, A. J. K., Eames, K. T. D, Gage, J. A., von Kirchbach, J. C., Ross, J. V., Saenz, R. A. and Gog, J. R. (2010). Measuring social networks in British primary schools through scientific engagement. *Proceedings of the Royal Society Series B* **278**, 1467–1475.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81**, 832–842.

GOMMANS, R. AND CILLESSEN, A. H. N. (2015). Nominating under constraints: a systematic comparison of unlimited and limited peer nomination methodologies in elementary school. *International Journal of Behavioral Development* **39**, 77–86.

GOODREAU, S. M. (2007). Advances in exponential random graph (p*) models applied to a large social network. *Social Networks* **29**, 231–248.

GOODREAU, S. M., KITTS, J. A. AND MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks*. *Demography* **46**, 103–125.

HANDCOCK, M. S. AND GILE, K. (2007). Modeling Social Networks with Sampled or Missing Data. Working Paper no. 75, Center for Statistics and the Social Sciences, University of Washington.

HIPP, J. R., WANG, C., BUTTS, C. T., JOSE, R. AND LAKON, C. M. (2015). Research note: the consequences of different methods for handling missing network data in stochastic actor based models. *Social Networks* **41**, 56–71.

HOFF, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory* **15**, 261–272.

HOFF, P. D., FOSDICK, B., VOLFOVSKY, A. AND STOVEL, K. (2013). Likelihoods for fixed rank nomination networks. *Network Science* **1**, 253–277.

HOFF, P., FOSDICK, B., VOLFOVSKY, A. AND HE, Y. (2015). *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*. R package version 1.1.

HOFF, P. D., RAFTERY, A. E. AND HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97**, 1090–1098.

HOLLAND, P. W. AND LEINHARDT, S. (1973). The structural implications of measurement error in sociometry. *Journal of Mathematical Sociology* **3**, 85–111.

KNOKE, D. AND YANG, S. (2008). *Social Network Analysis. Quantitative Applications in the Social Sciences*, 2nd edition, Volume 154. Thousand Oaks, CA: Sage Publications.

KOSSINETS, G. (2006). Effects of missing data in social networks. *Social Networks* **28**, 247–268.

MARSDEN, P. V. (1990). Network data and measurement. *Annual Review of Sociology* **16**, 435–463.

MERCKEN, L., SNIJDERS, T. A. B, STEGLICH, C., VARTIANEN, E. AND DE VRIES, H. (2010). Dynamics of adolescent friendship networks and smoking behavior. *Social Networks* **32**, 72–81.

MOSSONG, J., HENS, N., JIT, M., BEUTELS, M., AURANEN, P. AND MIKOLAJCZYK, K. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* **5**, e74.

POTTER, G. E., HANDCOCK, M. S., LONGINI, I. M. Jr and HALLORAN, M. E. (2011). Estimating within-household contact networks from egocentric data. *The Annals of Applied Statistics* **5**, 1816–1838.

RESNICK, M. D., BEARMAN, P. S., BLUM, R. W., BAUMAN, K. E., HARRIS, K. M., JONES, J., TABOR, J., BEUHRING, T., SIEVING, R. E., SHEW, M., IRELAND, M., BEARINGER, L. H. *and others*. (1997). Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association* **278**, 823–832.

ROBINS, G., PATTISON, P. AND WOOLCOCK, J. (2004). Missing data in networks: exponential random graph (p*) models for networks with non-respondents. *Social Networks* **26**, 257–283.

SMITH, J. A. AND MOODY, J. (2013). Structural effects of network sampling coverage I: nodes missing at random. *Social Networks* **35**, 652–668.

WANG, C., BUTTS, C. T., HIPP, J. R., JOSE, R. AND LAKON, C. M. (2016). Multiple imputation for missing edge data: a predictive evaluation method with application to add health. *Social Networks* **45**, 89–98.

WARNER, R. M., KENNEY, D. A. AND STOTO, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology* **37**, 1742–1757.

WITVLIET, M., BRENDGEN, M., van LIER, P., KOOT, H. M. and VITARO, F. (2010). Early adolescent depressive symptoms: prediction from clique isolation, loneliness, and perceived social acceptance. *Journal of Abnormal Child Psychology* **38**, 1045–1056.

YAN, B. AND GREGORY, S. (2011). Finding missing edges and communities in incomplete networks. *Journal of Physics A: Mathematical and Theoretical* **44**, 495102.

[*Received August 28, 2016; revised November 2, 2017; accepted for publication November 22, 2017*]